北京邮电大学实践课程实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程名称** | **创新创业实践** | **学院** | **计算机学院** | **指导老师** |
| **知识模块** | **大数据** | **完成时间** | | **2022年11月1日** |
| **班级** | **序号** | **学号** | **学生姓名** | **成绩** |
| **2022211313** | **5** | **2022211363** | **谢牧航** |  |

报告内容包括：

实验目的：

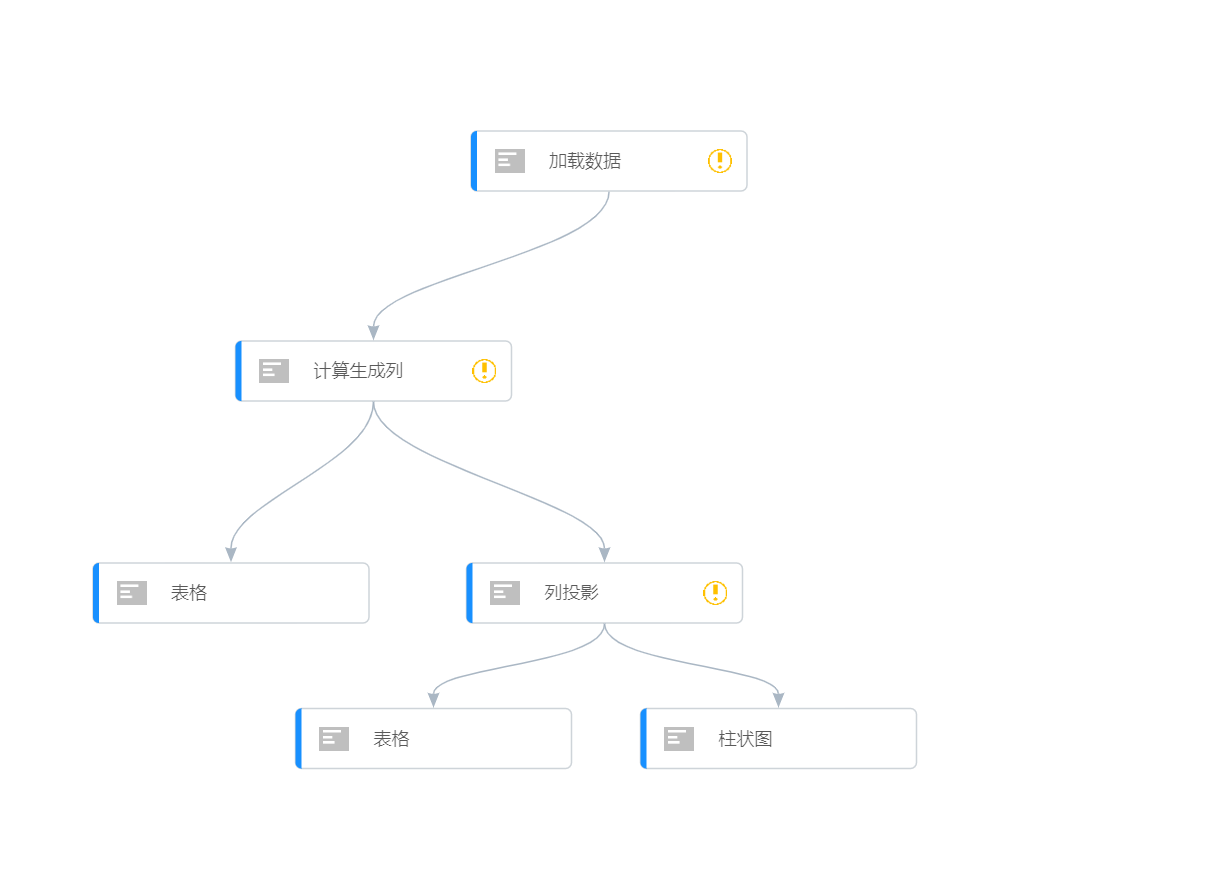
学习大数据基础功能，体验 BDAP 平台，进行一些大数据基础实验。

实验内容：

进行基础实验包括文件管理系统、数据挖掘、案例讲解、深度学习探索、在线编程等。设计工具包括数据预处理，决策树，K-means 聚类，SVM，Scikit-learn等。

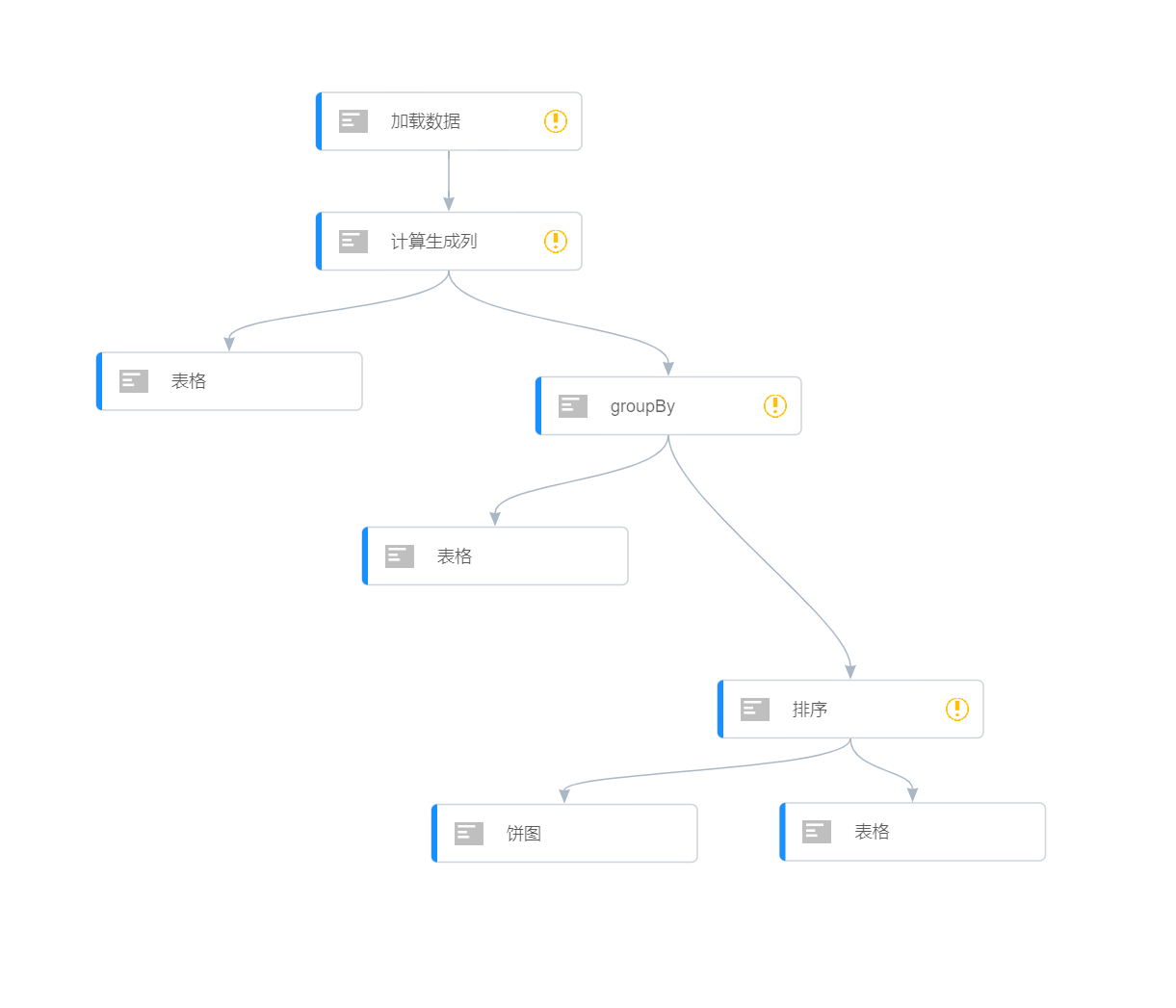
实验步骤：

**实验一：**熟悉大数据实验平台：进行最基本的工作流搭建和可视化等。



**实验二：**数据预处理

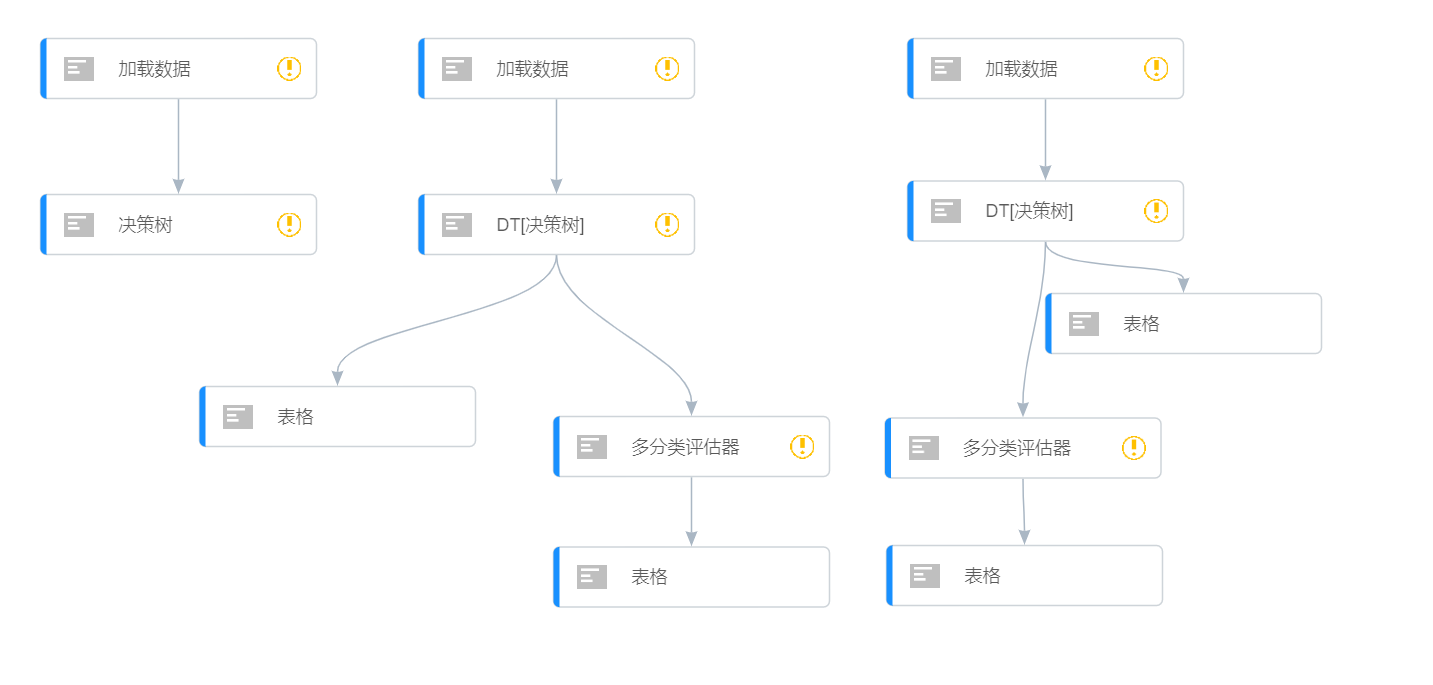
搭建工作流：



其中通过 groupby 根据一定的规则进行分组。通过一定的规则将一个数据集划分成若干个小的区域，然后针对若干个小区域进行数据处理。排序则对数据特定区域进行排序。

**实验三：**决策树模型

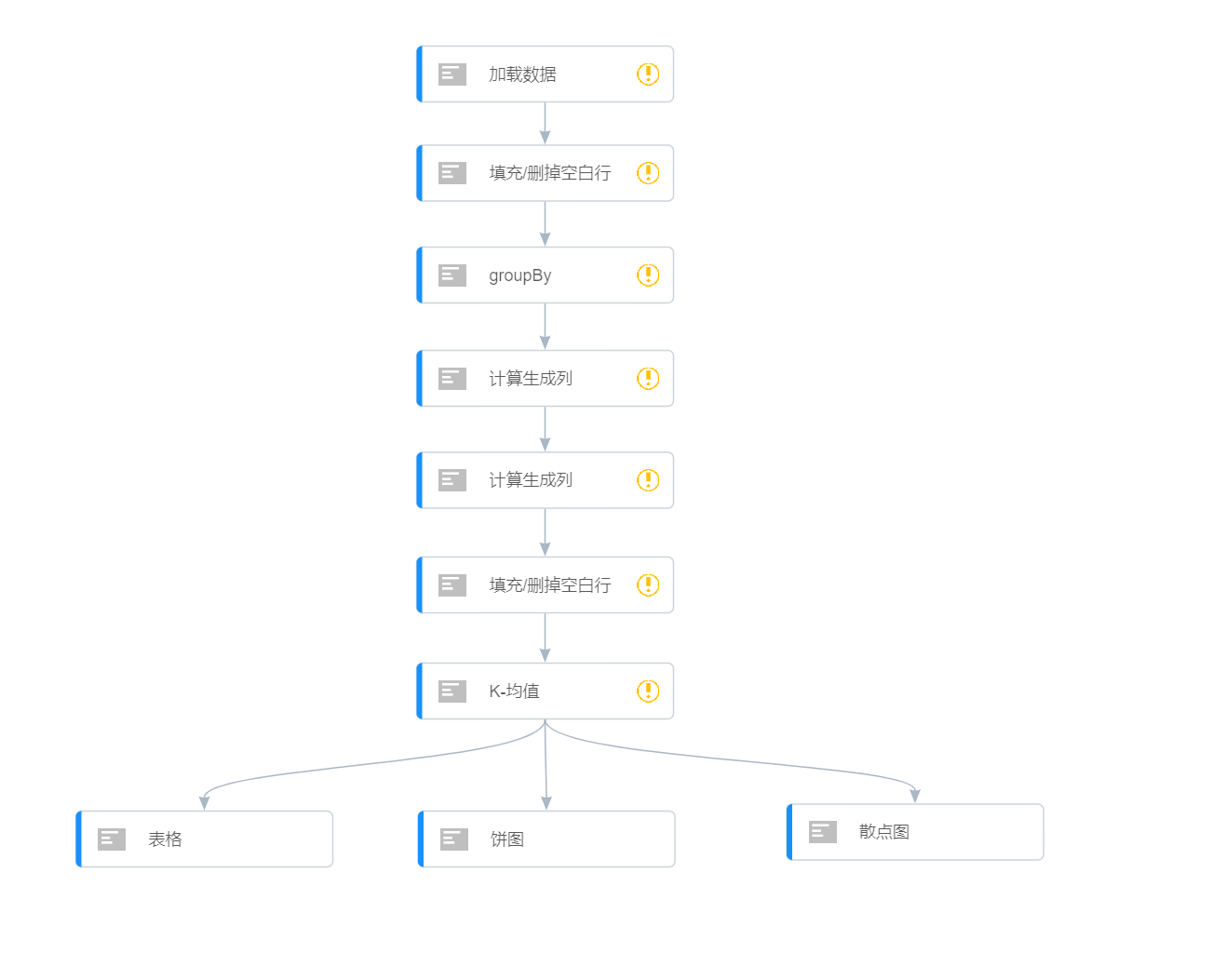
搭建工作流：



通过训练集对决策树进行训练，产生模型。并用评估集进行测试，并用多分类评估器对其 accuracy 进行评估。最后对测试集进行决策。

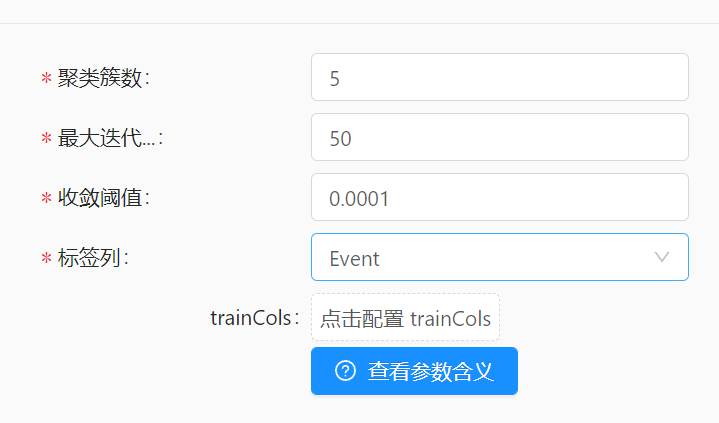
**实验四：**K-means 聚类

1：

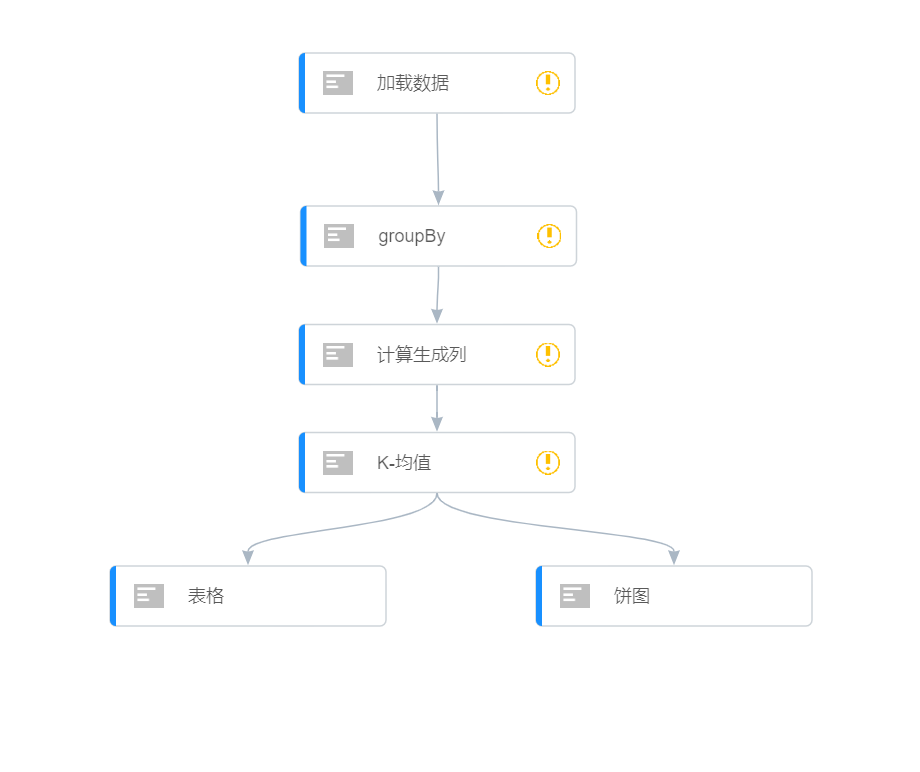


进行分组后再使用 K-means 聚类。

K-means 参数：

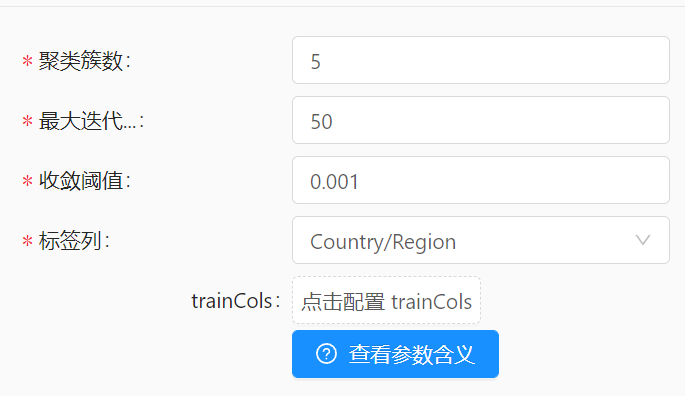


2.



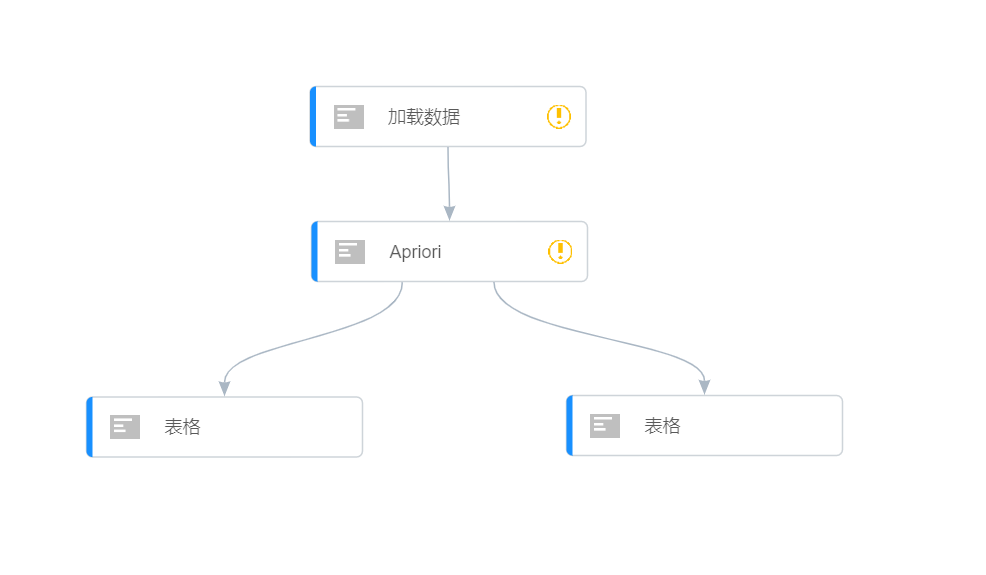
使用 K-means 探索新冠疫情数据和地区的关系。

K-means 参数



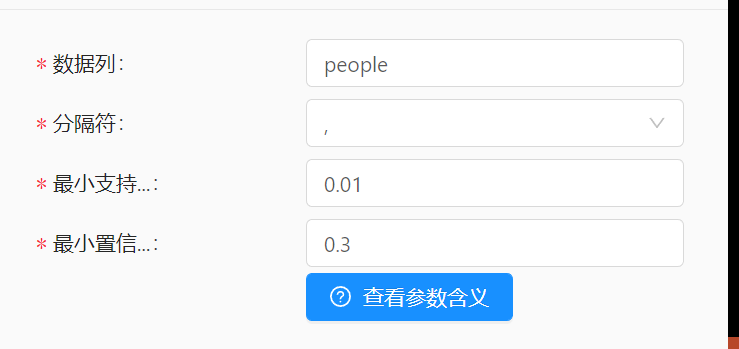
**实验五：**关联规则

搭建工作流：

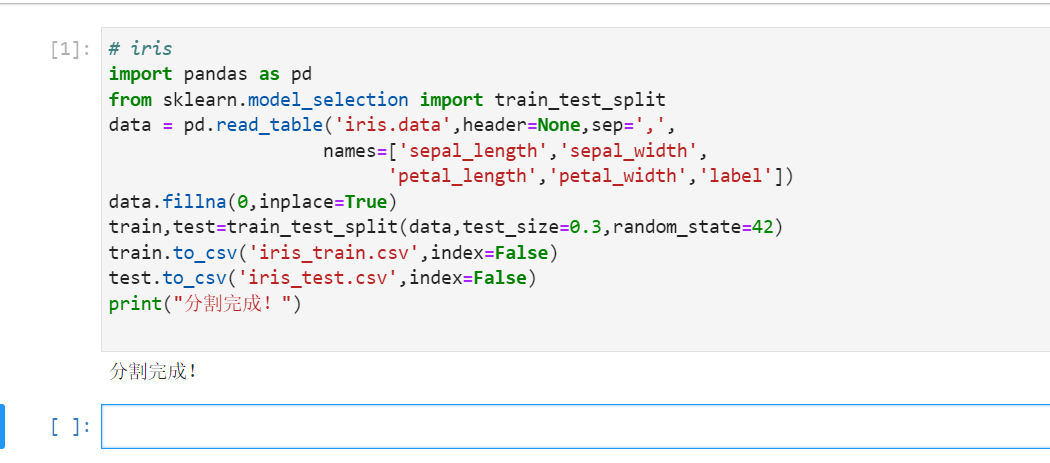


使用 Apriori 进行关联。

Apriori 参数：

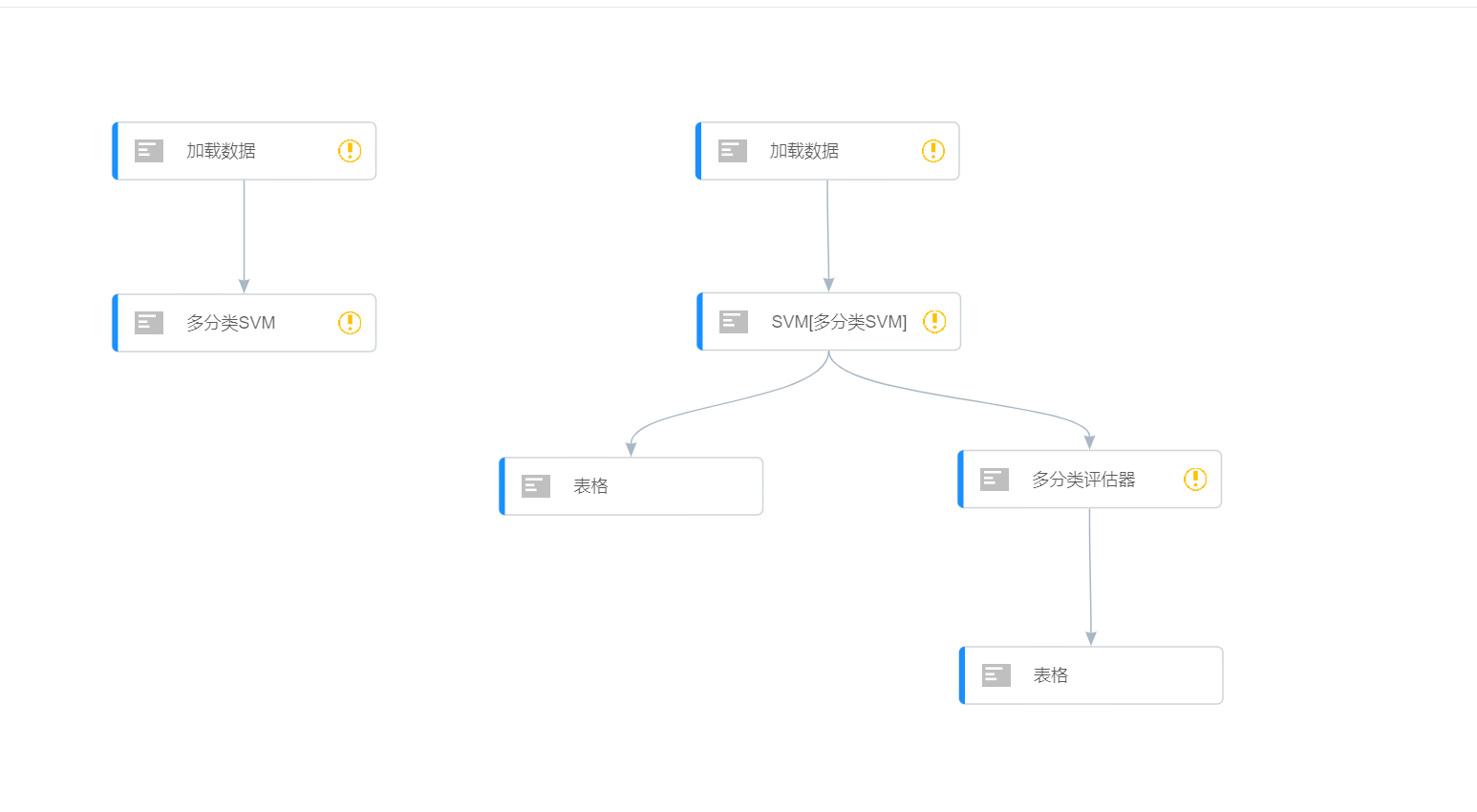


**实验六**：SVM 拓展实验

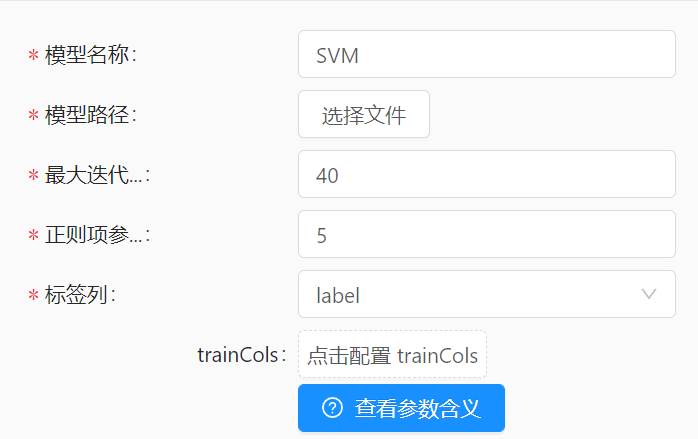
首先去 UCI 下载数据，并用 pandas 库进行数据格式转换，用sklearn库将数据集划分为训练集和测试集。

将分割好的 csv 文件上传到 BDAP 平台。

搭建工作流：



SVM 参数：



**实验七：**手写数字预测实验

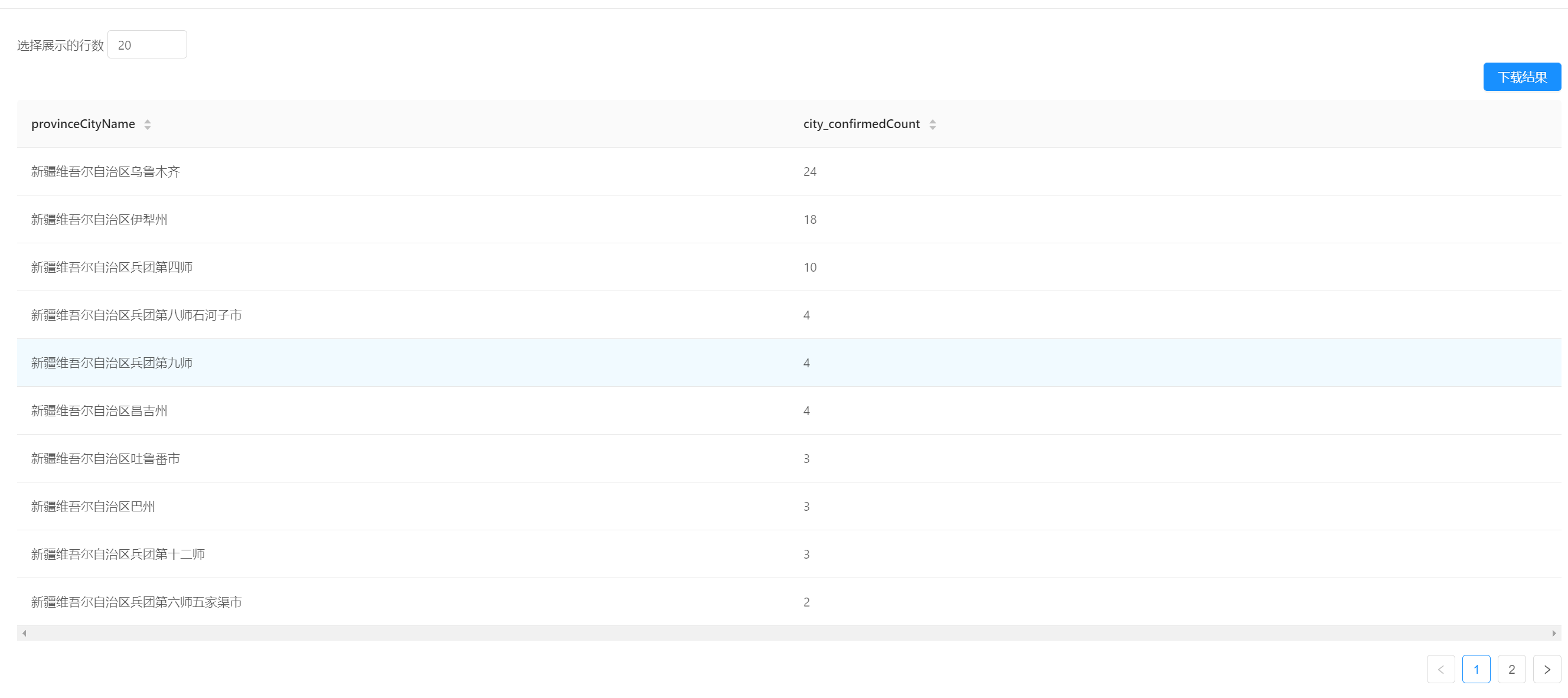
补全三个 TODO

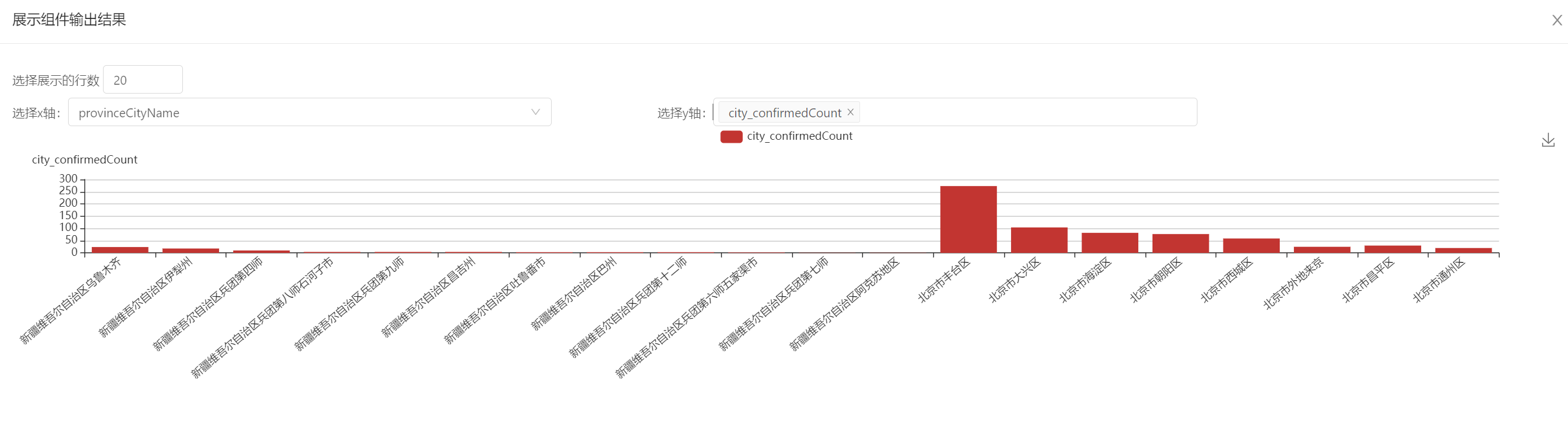


并在 Jupyter Notebook 中运行。

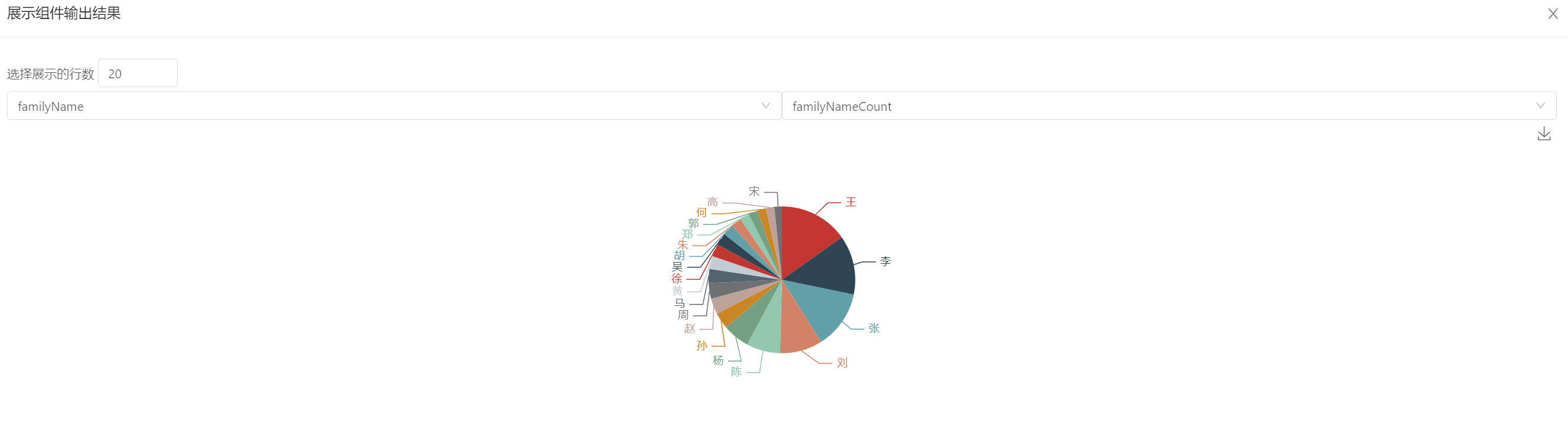
实验结果：

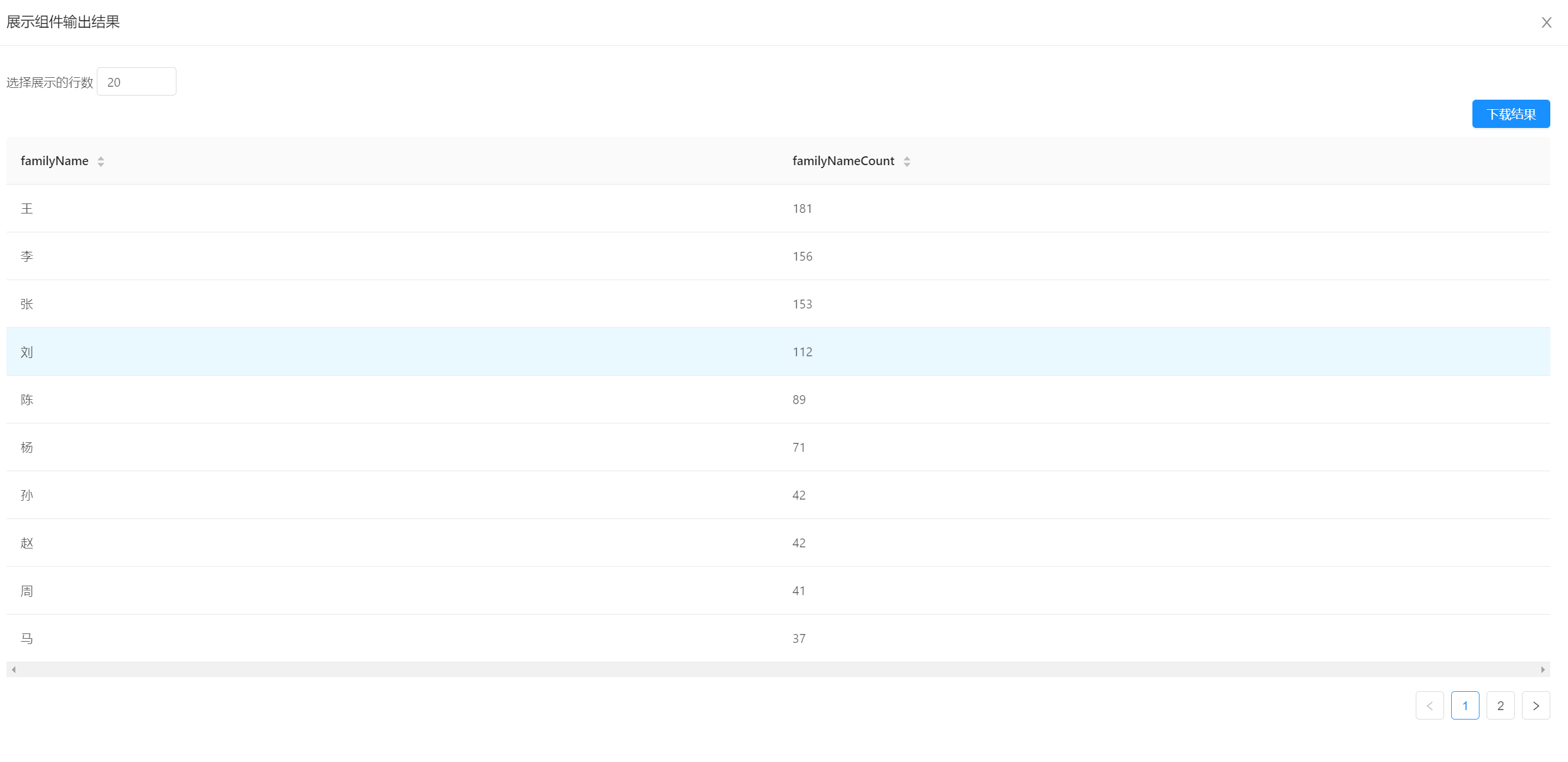
**实验一：**





**实验二：**



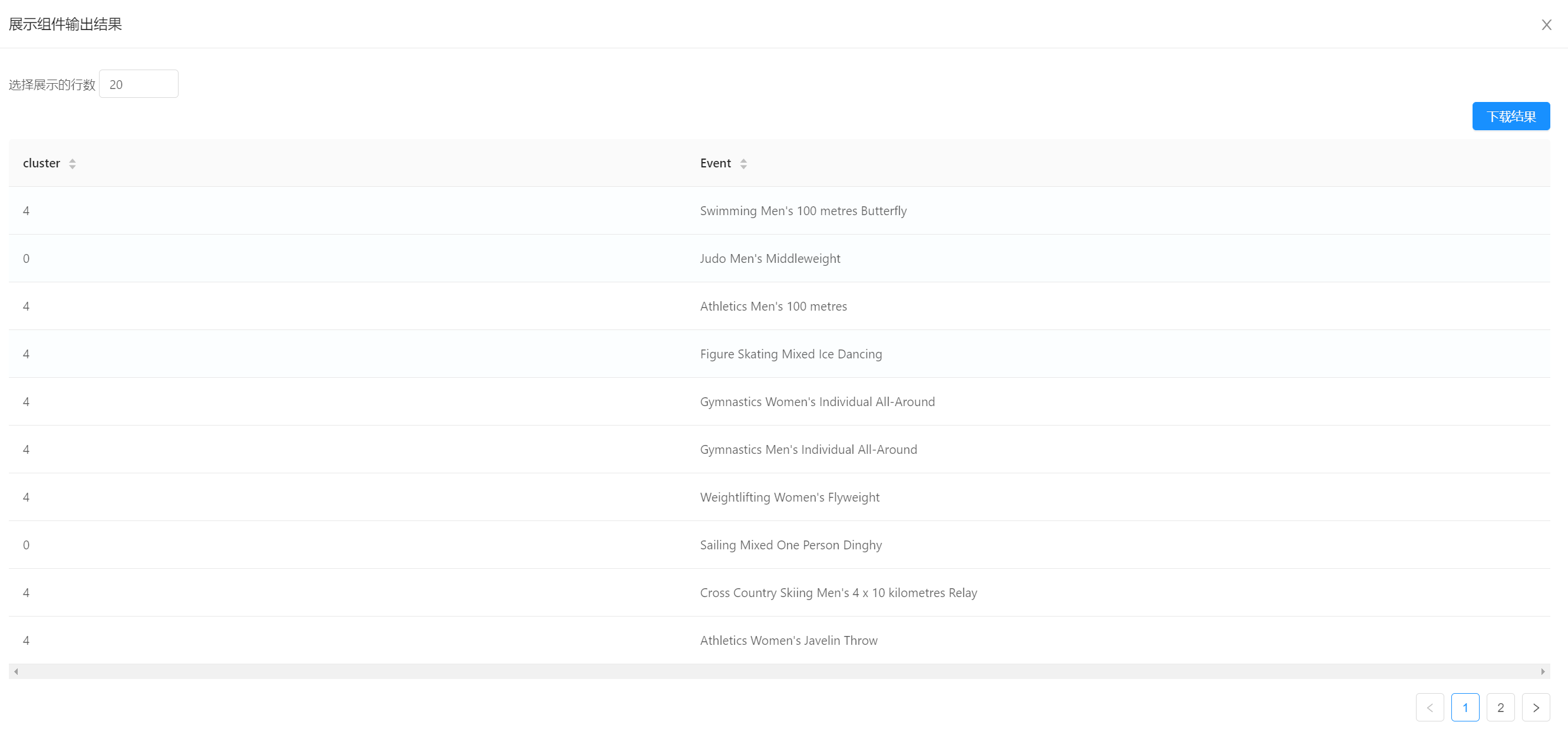


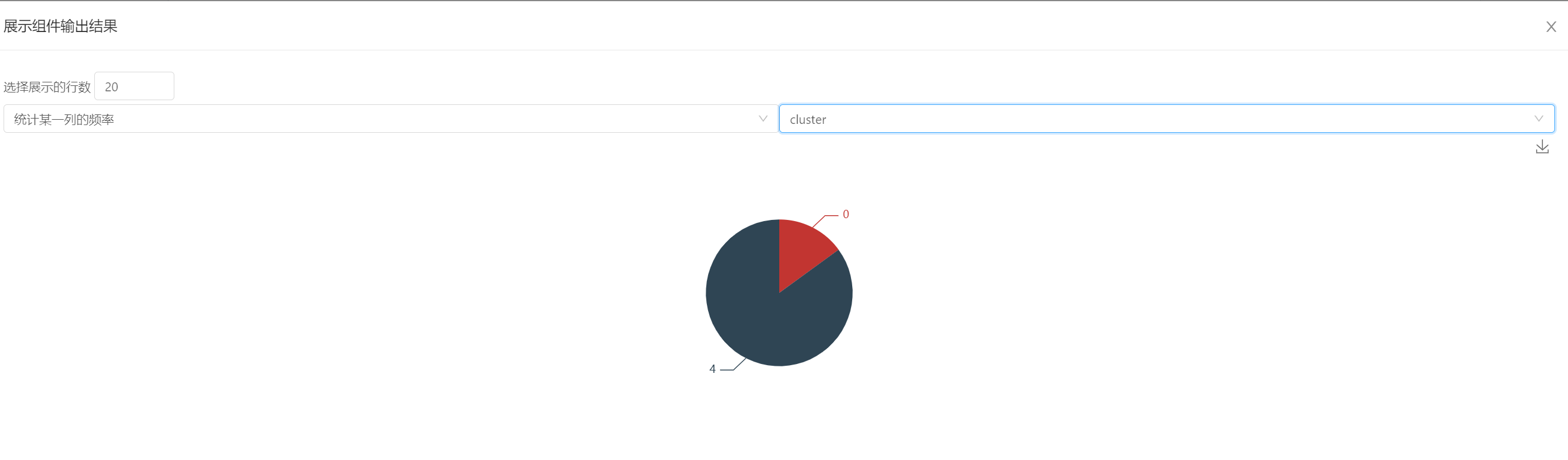
**实验三：**

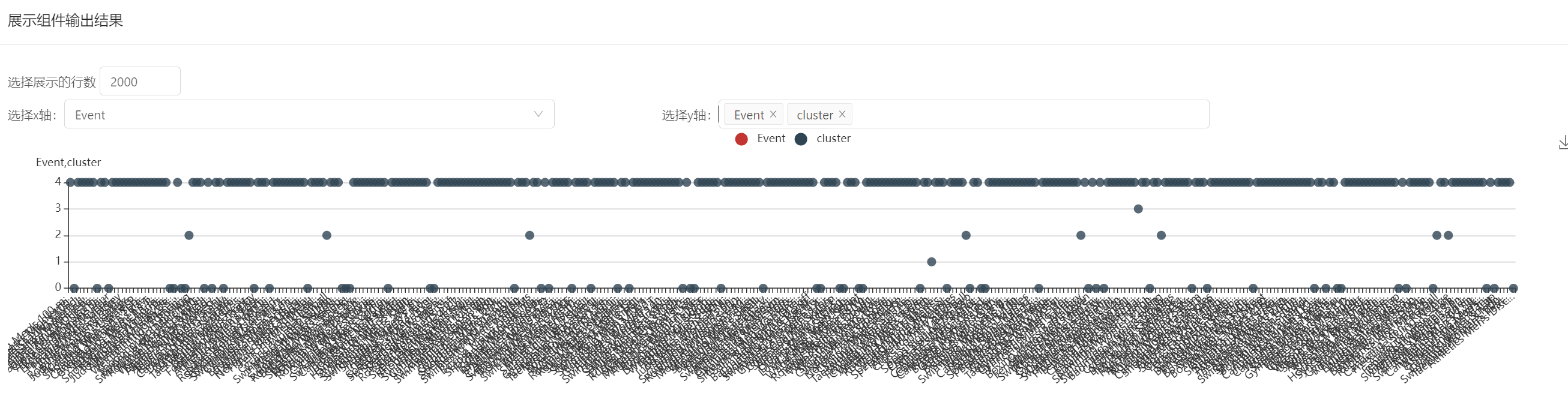


**实验四：**

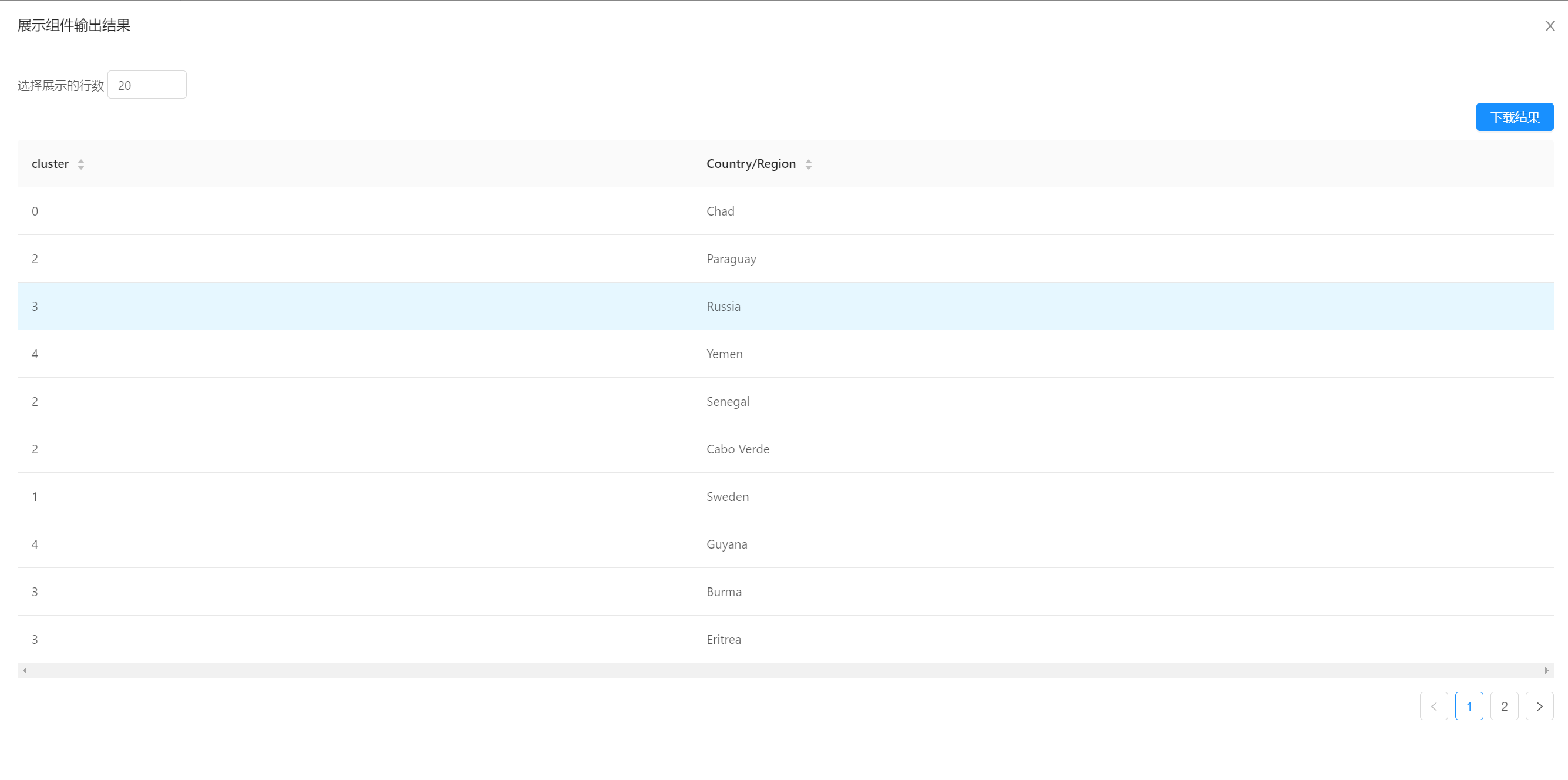
1.

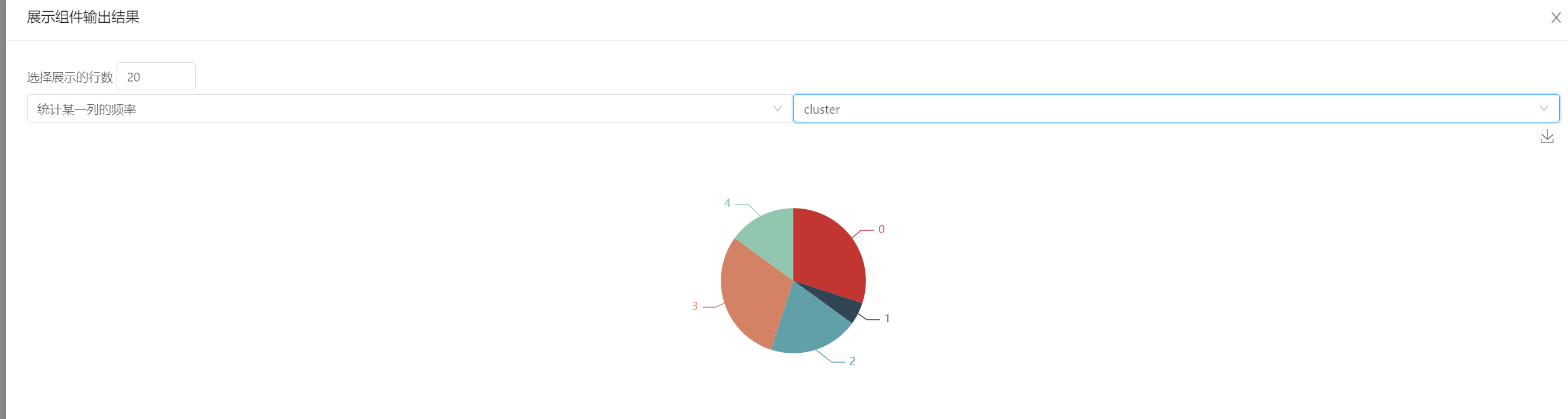




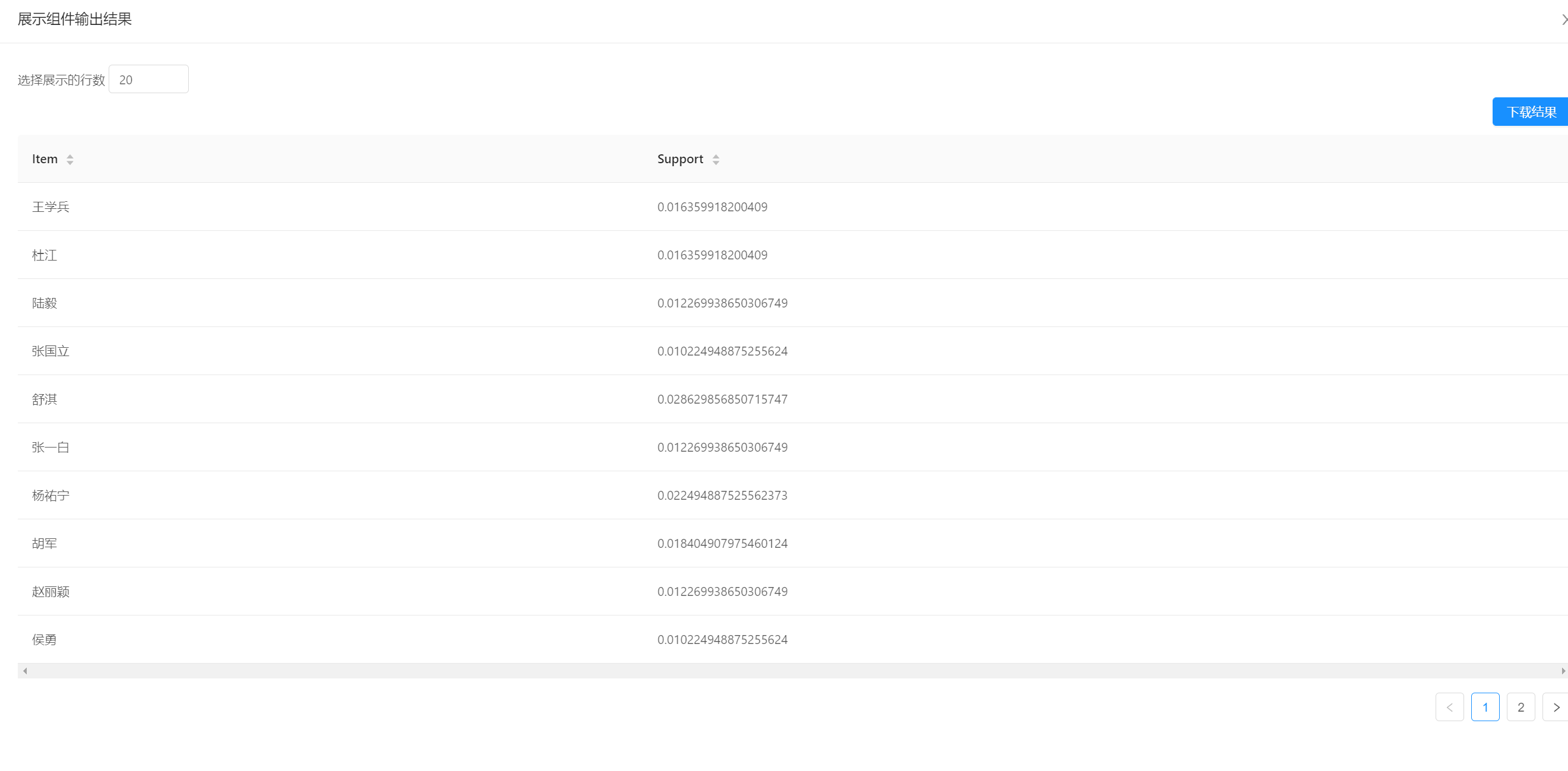


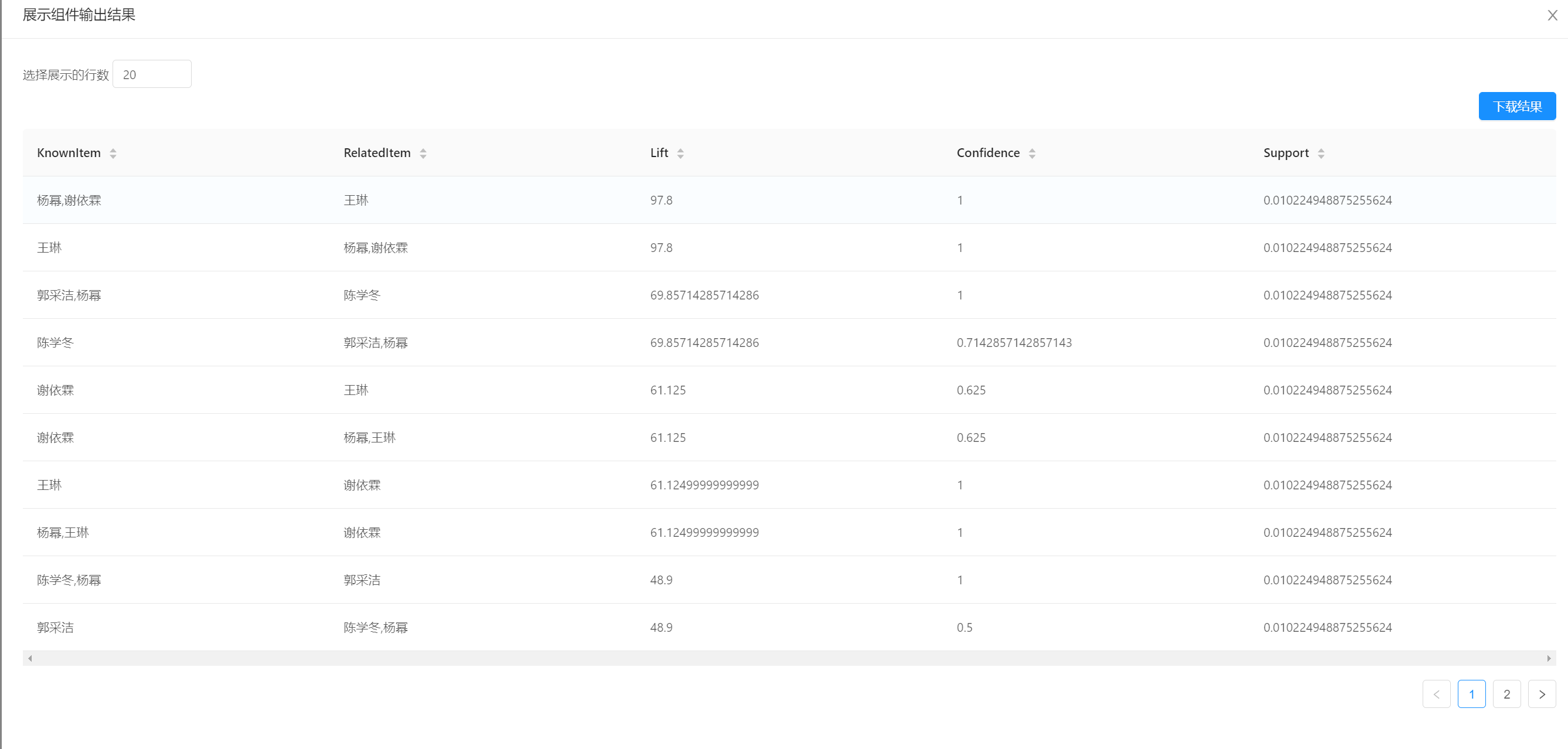
2.



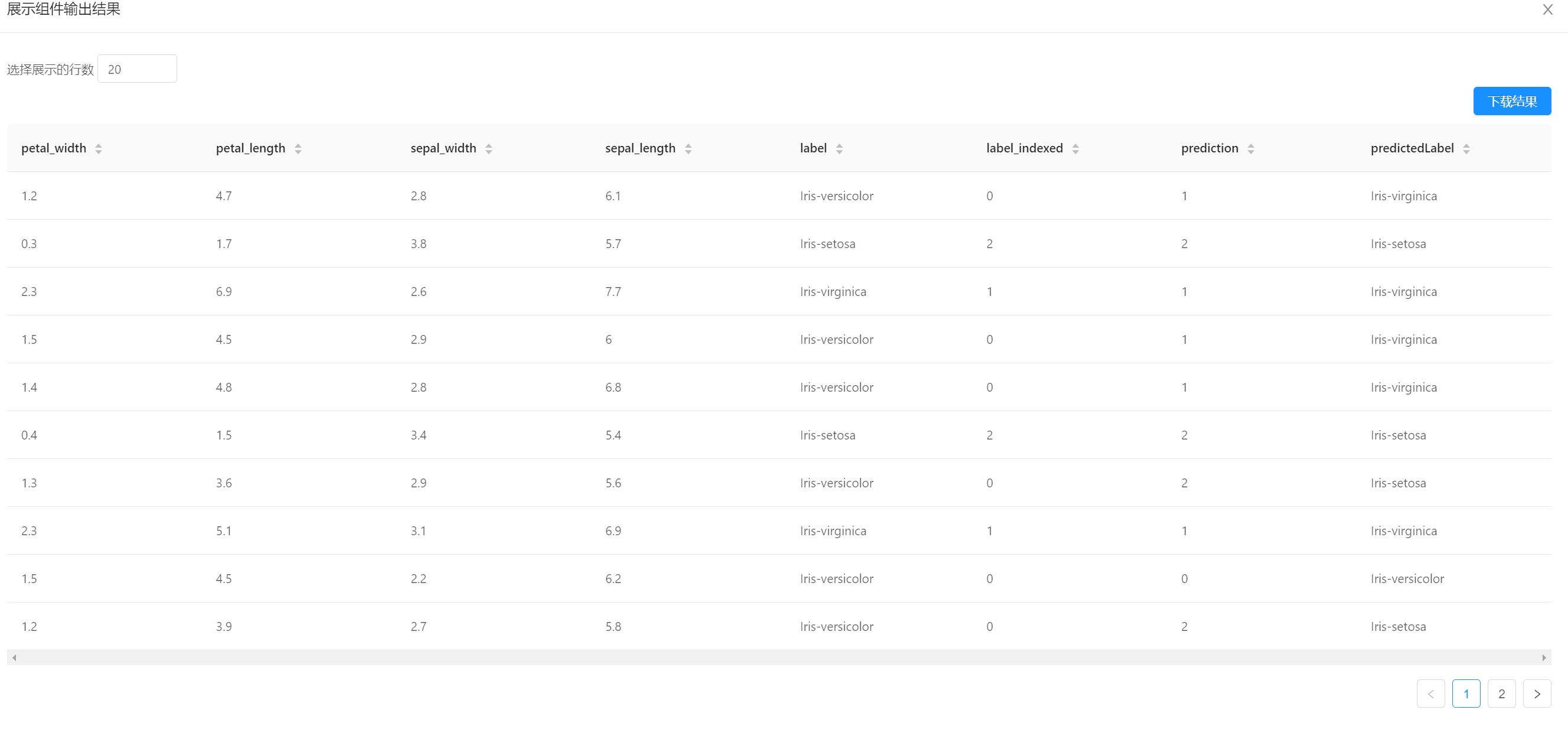


**实验五：**



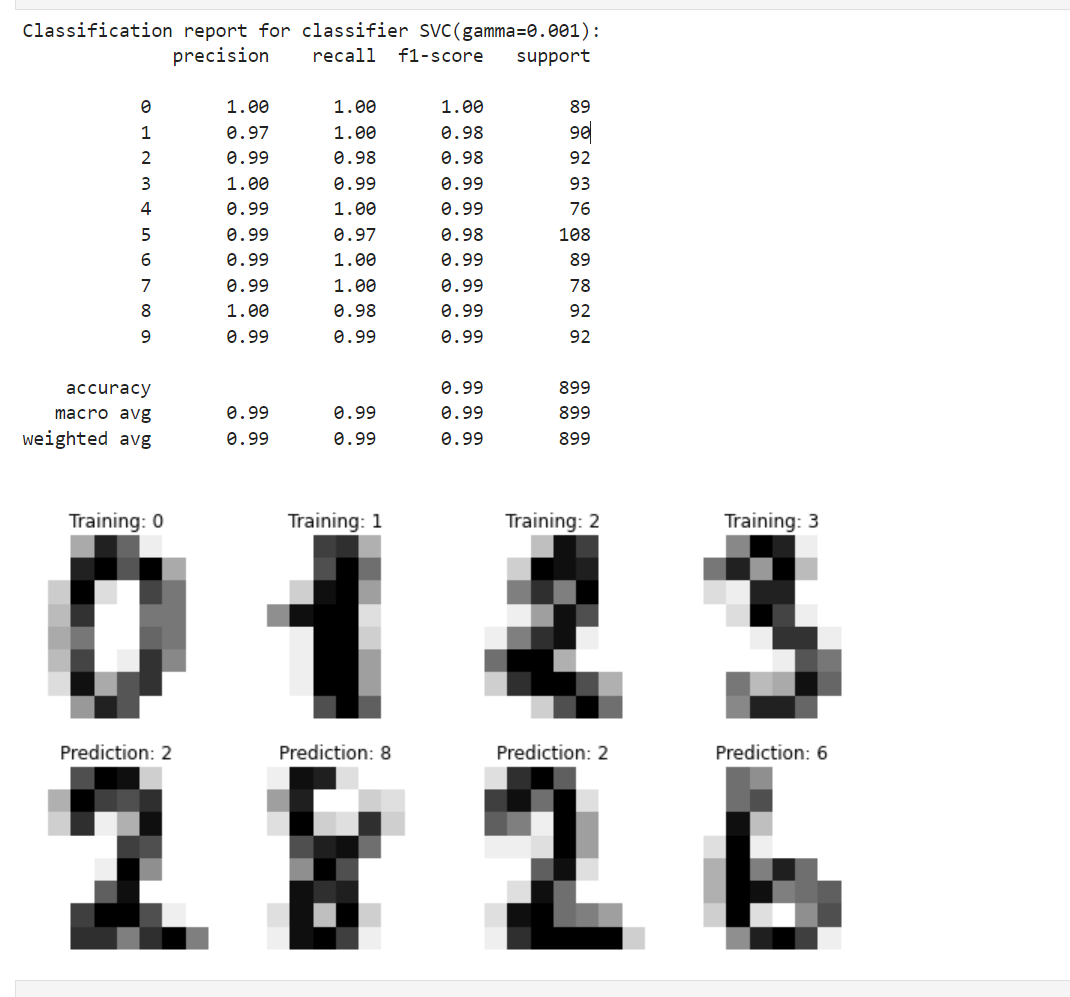


**实验六：**





**实验七：**



实验总结：

通过七个实验，对大数据有了基本的了解，也发现了其与之前机器学习的相关性。而且其中有很多自我探索的部分，做完之后能有更全面深刻的理解。并且对每个实验所用的工具的基本原理也有了了解。